Bayesian linear size-and-shape regression with applications to face data

Ian L. Dryden, Kwang-Rae Kim and Huiling Le,
School of Mathematical Sciences, University of Nottingham, UK.

Abstract

Regression models for size-and-shape analysis are developed, where the model is specified in the Euclidean space of the landmark coordinates. Statistical models in this space (which is known as the top space or ambient space) are often easier for practitioners to understand than alternative models in the quotient space of size-and-shapes. We consider a Bayesian linear size-and-shape regression model in which the response variable is given by labelled configuration matrix, and the covariates represent quantities such as gender and age. It is important to parameterize the model so that it is identifiable, and we use the LQ decomposition in the intercept term in the model for this purpose. Gamma priors for the inverse variance of the error term, matrix Fisher priors for the random rotation matrix, and flat priors for the regression coefficients are used. Markov chain Monte Carlo algorithms are used for sampling from the posterior distribution, in particular by using combinations of Metropolis-Hastings updates and a Gibbs sampler. The proposed Bayesian methodology is illustrated with an application to forensic facial data in three dimensions, where we investigate the main changes in growth by describing relative movements of landmarks for each gender over time.

1 Introduction

Bayesian linear regression analysis has been extensively studied for various types of response variables and covariates, where prior distributions are specified for the parameters in the classical regression model and statistical inference is carried out using the joint posterior distribution of the parameters (Gelman et al., 2013).

We wish to explore regression models for landmark data, where the location and orientation of the objects can be ignored. Such objects can be represented as points in the size-and-shape space (Dryden and Mardia, 2016, Chapter 5), which is defined as the space of landmark co-ordinates after rotation and translation information has been removed (Kendall, 1989). The shape space on the other hand (Dryden and Mardia, 2016, Chapter 4) is the space of landmark co-ordinates after rotation, translation *and scale* information has been removed (Kendall, 1986). Throughout this paper we will concentrate on size-and-shape rather than shape.

The size-and-shape space is a quotient space, where location and rotation are quotiented out by using least squares optimization. However, the geometry of the size-and-shape quotient space is complicated (Kendall et al., 1999), and it can be difficult for a practitioner to understand the meaning of statistical models formulated in the quotient space.

An alternative approach is to specify a statistical model in the Euclidean space of the landmark co-ordinates and then integrate out the unwanted location and rotation information by considering the marginal distribution of size-and-shape. In this case, the space in which the statistical model is specified is called the top space in differential geometry, and also known as the ambient space by some authors (Cheng et al., 2016). A top space modelling approach has the advantage that the model is often easier to understand than a quotient space model, and relatively standard inference methods can be used. We shall develop a Bayesian linear model in the space of the Euclidean landmark co-ordinates, and carry out statistical inference using Markov chain Monte Carlo (MCMC) algorithms. Care needs to be taken with identifiability of parameters in the model, and this issue often arises in high-dimensional object data (Dryden, 2014).

We consider a Bayesian regression model with response given by the size-and-shape of landmarks with real-valued covariates. A wide variety of regression problems on non-Euclidean spaces have been considered in previous work, and a summary of some approaches is given by Dryden and Mardia (2016, Section 13.4). Some approaches include directional data regression (Mardia, 1975; Mardia and Jupp, 2000; Presnell et al., 1998), tangent space regression models (Kent et al., 2001; Bowman, 2008; Faraway, 2004), growth curve models (Goodall and Lange, 1989), geodesic regression (Le and Kume, 2000; Hotz et al., 2010), principal geodesic analysis (Fletcher et al., 2004; Fletcher, 2013), geodesic PCA (Huckemann et al., 2010; Kenobi et al., 2010), principal nested spheres (Jung et al., 2012), intrinsic regression (Davis et al., 2007; Shi et al., 2009; Hinkle et al., 2014; Cornea et al., 2017), sphere-on-sphere regression (Rosenthal et al., 2014; Rosenthal et al., 2017; Di Marzio et al., 2018), unrolling and unwrapping (Jupp and Kent, 1987; Kume et al., 2007), manifold splines (Su et al., 2012) and many applications (e.g. Machado and Leite, 2006; Zhu et al., 2009; Samir et al., 2012; Yuan et al., 2012; Piras et al., 2014).

The remainder of this paper is organized as follows. In Section 2 we describe the Bayesian linear size-and-shape regression model, including the prior and posterior distributions. In Section 3, methods for Bayesian inference for the coefficients and model selection are presented. Finally an application to forensic facial data is given in Section 4.

2 Bayesian linear size-and-shape regression model

2.1 Linear model

Consider a random sample of n configurations of k labelled landmarks in m dimensions, where each configuration is represented by a $k \times m$ matrix $Y_i \in \mathbb{R}^{k \times m}, k > m, i = 1, \ldots, n$. We are interested only in the size-and-shapes of Y_i after removing translation and rotation, but preserving scale information (Dryden and Mardia, 2016, Chapter 5). In addition we have real valued covariates $x_{ij}, j = 1, \ldots, p$, corresponding to each configuration and without loss of generality we assume that each covariate is centred, i.e. $\sum_i x_{ij} = 0$. Categorical variables with g levels can be represented by $g - 1$ binary indicator variables in the standard way. We write $\mathbf{x}_i = (1, x_{1j}, \ldots, x_{pj})^\top$ as a $(p+1)$-dimensional column vector containing the p covariates and 1 for the intercept. We aim to predict the size-and-shape of Y_i using the covariates, and explore the relationship between Y_i and $\mathbf{x}_i$, $i = 1, \ldots, n$.

Suppose that Y_i are modelled with a probability distribution with conditional mean function $\mu(\mathbf{x}_i)$ given covariates $\mathbf{x}_i$ and subject to an arbitrary unknown rotation $\Lambda_i \in SO(m)$, where $SO(m)$ is the group of special orthogonal matrices that satisfy $\Lambda_i \Lambda_i^\top = \Lambda_i^\top \Lambda_i = I_m$ and $\det(\Lambda_i) = 1$, and where I_m

is the $m \times m$ identity matrix. So we have the conditional mean

$$E[Y_i|\mathbf{x}_i] = \mu(\mathbf{x}_i)\Lambda_i, \qquad i = 1, \ldots, n.$$

Including a noise term we have the model

$$Y_i = \mu(\mathbf{x}_i)\Lambda_i + \varepsilon_i,$$

where ε_i are assumed to be i.i.d. random matrix normal variables of dimension $k \times m$ (Gupta and Nagar, 1999, Chapter 2). In this paper we consider the conditional mean function $\mu(\mathbf{x}_i)$ to be linear so that the following linear regression model is of interest,

$$Y_i = \left(\alpha_0 + \sum_{j=1}^{p} \alpha_j x_{ij}\right)\Lambda_i + \varepsilon_i, \tag{1}$$

where $\alpha_0, \alpha_j \in \mathbb{R}^{k\times m}, j = 1, \ldots, p$, are $k \times m$ regression parameter matrices, the errors are matrix normal

$$\varepsilon_i \overset{i.i.d.}{\sim} MN_{k\times m}\big(\mathbf{0}, \sigma^2 I_m, I_k\big),$$

and so

$$\text{vec}(\varepsilon_i) \overset{i.i.d.}{\sim} N_{km}\big(\text{vec}(\mathbf{0}), \sigma^2 I_m \otimes I_k\big),$$

where $\text{vec}(A)$ denotes the vectorization of the matrix A (i.e. stacking columns) and $\otimes$ denotes the Kronecker product. The model (1) is not identifiable since the rotation effect from Λ_i dictates the coefficients $\{\alpha_0, \alpha_j\}$. We can make the model identifiable using an LQ decomposition of α_0. In particular we write $\alpha_0 = \beta_0 Q_0$, where $Q_0 \in SO(m)$ and β_0 is lower triangular (i.e. has zero entries above the leading diagonal). Therefore the model (1) can be rewritten as

$$\begin{aligned} Y_i &= \mu(\mathbf{x}_i)\Gamma_i + \varepsilon_i, \\ &= \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)\Gamma_i + \varepsilon_i \\ &= X_i\beta\Gamma_i + \varepsilon_i, \end{aligned} \tag{2}$$

where $X_i = \mathbf{x}_i^\top \otimes I_k \in \mathbb{R}^{k\times k(p+1)}$ is a $k \times k(p+1)$ matrix, $\beta = \begin{bmatrix}\beta_0^\top & \beta_1^\top & \cdots & \beta_p^\top\end{bmatrix}^\top \in \mathbb{R}^{k(p+1)\times m}$ is a $k(p+1) \times m$ matrix of regression parameters and $\Gamma_i \in SO(m)$. In the following we describe a Bayesian approach to estimate $\mu(\mathbf{x}_i)$ given deterministic covariates $\mathbf{x}_i$.

2.2 Likelihood

It follows from the matrix normality of ε_i that $Y_i \sim MN_{k\times m}\big(X_i\beta\Gamma_i, \sigma^2 I_m, I_k\big)$, therefore the probability density function of Y_i is given by

$$f(Y_i \mid \beta, \Gamma_i, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{km/2}} \exp\left(-\frac{1}{2\sigma^2}\,\text{tr}\left[(Y_i - X_i\beta\Gamma_i)^\top(Y_i - X_i\beta\Gamma_i)\right]\right)$$

and the likelihood is given by

$$f(Y_1, \ldots, Y_n, \mid \beta, \Gamma_1, \ldots, \Gamma_n, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{nkm/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\text{tr}\left[(Y_i - X_i\beta\Gamma_i)^\top(Y_i - X_i\beta\Gamma_i)\right]\right).$$

2.3 Prior and posterior

We shall concentrate on the $m = 3$ dimensional case and it is then helpful to adopt a particular parameterization of the rotation matrices. We can represent the three dimensional rotation matrix using the ZXZ-convention where

$$\Gamma(\theta_1, \theta_2, \theta_3) = \begin{bmatrix} \cos\theta_3 & \sin\theta_3 & 0 \\ -\sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_2 & \sin\theta_2 \\ 0 & -\sin\theta_2 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} \cos\theta_1 & \sin\theta_1 & 0 \\ -\sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$0 \leq \theta_1, \theta_3 < 2\pi$ and $0 \leq \theta_2 < \pi$ (Landau and Lifschitz, 1976). If we assume that $\theta_1, \theta_3 \sim U[0, 2\pi)$ and $\theta_2 \sim U[0, \pi)$, then using these co-ordinates the density of the uniform distribution on $SO(m)$ is

$$g(\theta_1, \theta_2, \theta_3) = \frac{1}{2\pi}\left(\frac{1}{2}\sin\theta_2\right)\frac{1}{2\pi} \propto \sin\theta_2. \tag{3}$$

We consider the following priors for parameters $(\kappa, \Gamma_i, \beta)$. Let $\kappa = 1/\sigma^2$. Assume that κ follows a Gamma distribution with shape parameter a and scale parameter b. We consider the prior for the rotation matrix to be the matrix Fisher distribution (Mardia and Jupp, 2000, p.89) and F_0 is a 3×3 parameter matrix of that so that $p(\Gamma_i; F_0) \propto \exp\{\mathrm{tr}(F_0^\top \Gamma_i)\}\sin\theta_{i2}$ and $\sin\theta_{i2}$ is due to the uniform measure. The regression parameters β are taken to be uniform and all the parameters are independent, i.e.

$$\begin{aligned} &\kappa \sim \text{Gamma}(a, b)\ ; \\ &\Gamma_i \sim \text{matrix Fisher}(F_0), \quad i = 1, \ldots, n\ ; \\ &p(\beta \mid \Gamma_1, \ldots, \Gamma_n, \kappa) \propto 1, \end{aligned}$$

independently. Then the joint posterior density for $(\beta, \Gamma_1, \ldots, \Gamma_n, \kappa)$ is given by

$$\begin{aligned} &p(\beta, \Gamma_1, \ldots, \Gamma_n, \kappa \mid Y_1, \ldots, Y_n) \\ \propto\ &\exp\left(\sum_{i=1}^n \mathrm{tr}(F_0^\top \Gamma_i)\right)\left[\prod_{i=1}^n \sin\theta_{i2}\right]\kappa^{a+3nk/2-1}\exp\left(-\frac{\kappa}{b}\right)\exp\left(-\frac{1}{2}\kappa\sum_{i=1}^n \mathrm{tr}\left[(Y_i - X_i\beta\Gamma_i)^\top(Y_i - X_i\beta\Gamma_i)\right]\right). \end{aligned}$$

The conditional posterior for $(\kappa \mid \Gamma_1, \ldots, \Gamma_n, \beta, Y_1, \ldots, Y_n)$ is

$$\kappa \mid \Gamma_1, \ldots, \Gamma_n, \beta, Y_1, \ldots, Y_n \sim \text{Gam}\left(a + \frac{3nk}{2}, \frac{1}{\frac{1}{b} + \frac{1}{2}\sum_{i=1}^n \mathrm{tr}\left[\left(Y_i - X_i\beta\Gamma_i\right)^\top\left(Y_i - X_i\beta\Gamma_i\right)\right]}\right).$$

The conditional posterior for $(\beta \mid \Gamma_1, \ldots, \Gamma_n, \kappa, Y_1, \ldots, Y_n)$ is

$$\mathrm{vec}(\beta^\top)^{(-0)} \mid \Gamma_1, \ldots, \Gamma_n, \kappa, Y_1, \ldots, Y_n \sim N_{3k(p+1)-3}\left(\mathrm{vec}(\xi^\top)^{(-0)}, (\Omega \otimes \Sigma)^{(-0)}\right),$$

where

$$\begin{aligned} \Sigma &= \frac{1}{\kappa} I_3, \\ \Omega &= \left(\sum_{i=1}^n X_i^\top X_i\right)^{-1}, \\ \xi &= \left(\sum_{i=1}^n X_i^\top X_i\right)^{-1}\sum_{i=1}^n X_i^\top Y_i \Gamma_i^\top, \end{aligned}$$

and

$$\text{vec}(\beta^\top) = \begin{bmatrix} \text{vec}(\beta_0^\top) \\ \text{vec}(\beta_1^\top) \\ \vdots \\ \text{vec}(\beta_p^\top) \end{bmatrix}, \quad \text{vec}(\xi^\top) = \begin{bmatrix} \text{vec}(\xi_0^\top) \\ \text{vec}(\xi_1^\top) \\ \vdots \\ \text{vec}(\xi_p^\top) \end{bmatrix} \quad \text{with size} \quad \begin{bmatrix} 3k \times 1 \\ 3k \times 1 \\ \vdots \\ 3k \times 1 \end{bmatrix},$$

$\Omega \otimes \Sigma$ is a $3k(p+1) \times 3k(p+1)$ covariance matrix, and $^{(-0)}$ stands for removing 2th, 3th, 6th elements of $\text{vec}(\beta^\top)$ and $\text{vec}(\xi^\top)$, and also removing those three rows and columns of $\Omega \otimes \Sigma$. Hence for each $\text{vec}(\beta_0^\top)^{(-0)}, \text{vec}(\beta_1^\top), \ldots, \text{vec}(\beta_p^\top)$ of length $3k-3, 3k, \ldots, 3k$, we can use the following conditional distribution of partitioned multivariate normal distribution

$$\begin{aligned} &\text{vec}(\beta_0^\top)^{(-0)} \mid \Gamma_1, \ldots, \Gamma_n, \kappa, Y_1, \ldots, Y_n, \text{vec}(\beta_1^\top), \ldots, \text{vec}(\beta_p^\top), \\ &\text{vec}(\beta_j^\top) \mid \Gamma_1, \ldots, \Gamma_n, \kappa, Y_1, \ldots, Y_n, \text{vec}(\beta_0^\top)^{(-0)}, \text{vec}(\beta_{-j}^\top), \quad j = 1, \ldots, p. \end{aligned}$$

The conditional posterior for $(\Gamma_1, \ldots, \Gamma_n \mid \kappa, \beta, Y_1, \ldots, Y_n)$ is proportional to

$$\exp\left(\sum_{i=1}^n \text{tr}\left[(F_0 + \kappa\beta^\top X_i^\top Y_i)^\top \Gamma_i\right]\right)\left[\prod_{i=1}^n \sin\theta_{i2}\right].$$

Hence for a specific ith observation, using independence the conditional posterior for Γ_i is proportional to

$$\Gamma_i \mid \Gamma_{-i}, \kappa, \beta, Y_1, \ldots, Y_n \propto \exp\left(\text{tr}\left[F_i^\top \Gamma_i\right]\right) \sin\theta_{i2},$$

where

$$F_i = F_0 + \kappa\beta^\top X_i^\top Y_i, \qquad i = 1, \ldots, n.$$

Let us drop the observation index i for a moment then

$$\begin{aligned} \text{tr}(F^\top\Gamma) &= C_1\cos\theta_1 + S_1\sin\theta_1 + R_1 \\ &= C_2\cos\theta_2 + S_2\sin\theta_2 + R_2 \\ &= C_3\cos\theta_3 + S_3\sin\theta_3 + R_3, \end{aligned}$$

where

$$\begin{aligned} C_1 &= F_{11}\cos\theta_3 - F_{21}\sin\theta_3 + F_{12}\sin\theta_3\cos\theta_2 + F_{22}\cos\theta_3\cos\theta_2 - F_{32}\sin\theta_2, \\ S_1 &= -F_{11}\sin\theta_3\cos\theta_2 - F_{21}\cos\theta_3\cos\theta_2 + F_{31}\sin\theta_2 + F_{12}\cos\theta_3 - F_{22}\sin\theta_3, \\ C_2 &= -F_{11}\sin\theta_3\sin\theta_1 - F_{21}\cos\theta_3\sin\theta_1 + F_{12}\sin\theta_3\cos\theta_1 + F_{22}\cos\theta_3\cos\theta_1 + F_{33}, \\ S_2 &= F_{31}\sin\theta_1 + F_{13}\sin\theta_3 + F_{23}\cos\theta_3 - F_{32}\cos\theta_1, \\ C_3 &= F_{11}\cos\theta_1 - F_{21}\cos\theta_2\sin\theta_1 + F_{12}\sin\theta_1 + F_{22}\cos\theta_2\cos\theta_1 + F_{23}\sin\theta_2, \\ S_3 &= -F_{11}\cos\theta_2\sin\theta_1 - F_{21}\cos\theta_1 + F_{12}\cos\theta_2\cos\theta_1 - F_{22}\sin\theta_1 + F_{13}\sin\theta_2, \end{aligned}$$

and R_1, R_2, R_3 are remainder terms independent of each Euler angle. Hence for ith observation, the conditional distributions for θ_{i1} and θ_{i3} are von Mises distributions (Green and Mardia, 2006). Since

$$\theta_{i2} \mid \theta_{i1}, \theta_{i3}, \Gamma_{-i}, \kappa, \beta, Y_1, \ldots, Y_n \propto \exp\left(C_{i2}\cos\theta_{i2} + S_{i2}\sin\theta_{i2}\right)\sin\theta_{i2},$$

we use a Metropolis-Hastings update for θ_{i2}. Hence for posterior sampling by MCMC we use Gibbs sampler for $(\kappa, \beta, \theta_{i1}, \theta_{i3})$, and Metropolis-Hastings algorithm for θ_{i2}.

Remark (Helmertized size-and-shape. Let $Y_i^H = HY_i$ be Helmertized size-and-shapes, where H is the Helmert sub-matrix (Dryden and Mardia, 2016, p.49-50). It is often useful to work with Y_i^H as this takes care of the location invariance for size-and-shapes, and reduces the number of parameters appropriately.

Condition (Identifiability). Consider the Helmertized model

$$Y_i^H = \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right) \Gamma_i + \varepsilon_i,$$

where $\beta_j, j = 0, 1, \ldots, p$, are $(k-1) \times 3$ matrices. Let G be the number of distinct sets of covariate tuples in $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, then for $k \geq 4$

$$p_1 = G\{3(k-1) - 3\} = G(3k-6)$$

is the number of regression parameters that can be identifiable. Let

$$p_2 = \{3(k-1) - 3\} + p\{3(k-1)\} = (3k-6) + p(3k-3) = 3k(p+1) - 6 - 3p$$

be the number of parameters in regression model. Then all p_2 parameters in regression model are identifiable if $p_1 \geq p_2$.

This identifiability condition indicates that the stability of estimation depends on how many distinct tuples of covariates are used. If $p_1 < p_2$ then MCMC draws of parameters can be away from the true values due to non-identifiability, or we may need a long number iterations if $p_1 = p_2$.

3 Inference and model selection

3.1 Inference for the coefficients

After the posterior sample $\{\beta_j^{(t)}, j = 0, \ldots, p, t = 1, \ldots, T\}$ is obtained from T iterations of the MCMC algorithm after burn-in, we can make an inference for β. Marginal $100(1-\alpha)\%$ credible intervals for $\beta_j, j = 0, \ldots, p$, are given by

$$\left[\beta_{j,\alpha/2}, \quad \beta_{j,1-\alpha/2}\right],$$

where $\beta_{j,\mathcal{P}}$ denotes the quantile at probability $\mathcal{P}$ based on order statistics from the sample after burn-in. Since $\beta_j^{(t)}$ is a matrix we define the matrix quantile as an element-wise quantile.

An alternative approach based on marginal Gaussian distributions for β_j is

$$\left[\widehat{\beta}_j - z_{\alpha/2} \cdot \widehat{sd(\beta_j)}, \quad \widehat{\beta}_j + z_{\alpha/2} \cdot \widehat{sd(\beta_j)}\right],$$

where

$$\begin{aligned} \widehat{\beta}_j &= \frac{1}{T}\sum_{t=1}^{T} \beta_j^{(t)}, \\ \widehat{sd(\beta_j)} &= \sqrt{\frac{1}{T-1}\sum_{t=1}^{T} \left(\beta_j^{(t)} - \widehat{\beta}_j\right) \circ \left(\beta_j^{(t)} - \widehat{\beta}_j\right)}, \end{aligned}$$

and $\circ$ is the Hadamard product defined by $(X \circ Y)_{i,j} = (X)_{i,j} \cdot (Y)_{i,j}$ for two matrices X and Y of the same dimension.

3.2 Model selection

We now present measures for model selection. For convenience write $\Theta = (\beta, \Gamma_1, \ldots, \Gamma_n, \sigma^2)$ and let $L(Y \mid \Theta)$ be the likelihood function. Define the deviance as $D(\Theta) = -2\log\big(L(Y \mid \Theta)\big)$, then the deviance information criterion (DIC) is defined by penalizing the deviance by the effective number of parameters, p_D (Spiegelhalter et al., 2002)(Gelman et al., 2013, p.172), i.e.

$$\begin{aligned} \text{DIC} &= \overline{D} + p_D \\ &= D(\overline{\Theta}) + 2p_D, \end{aligned}$$

where $\overline{D} = \mathbb{E}[D(\Theta)]$ is the posterior expected deviance and $\overline{\Theta}$ is the posterior mean of Θ. In practice the posterior expectations are obtained from the arithmetic means of the relevant terms from a MCMC algorithm after burn-in. The effective number of parameters can be estimated by either $p_D^{(1)} = \overline{D} - D(\overline{\Theta})$ (Spiegelhalter et al., 2002) or $p_D^{(2)} = \frac{1}{2}\text{var}\big(D(\Theta)\big)$ (Gelman et al., 2013, p.173).

The Watanabe-Akaike information criterion or widely available information criterion (WAIC) (Watanabe, 2010) (Gelman et al., 2013, p.173) is a fully Bayesian criterion based on the log pointwise posterior predictive density adjusted by the effective number of parameters, p_{WAIC}, to avoid overfitting and is defined by

$$\text{WAIC} = -2\left\{\sum_{i=1}^{n} \log\big(\mathbb{E}\big[f\big(Y_i \mid \beta, \Gamma_i, \sigma^2\big)\big]\big) + p_{\text{WAIC}}\right\},$$

where again we have two possible estimates of the effective number of parameters:

$$\begin{aligned} p_{\text{WAIC1}} &= 2\sum_{i=1}^{n}\Big(\log\big(\mathbb{E}\big[f\big(Y_i \mid \beta, \Gamma_i, \sigma^2\big)\big]\big) - \mathbb{E}\big[\log f\big(Y_i \mid \beta, \Gamma_i, \sigma^2\big)\big]\Big), \\ p_{\text{WAIC2}} &= \sum_{i=1}^{n}\text{var}\Big(\log f\big(Y_i \mid \beta, \Gamma_i, \sigma^2\big)\Big). \end{aligned}$$

The Akaike information criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) (Schwarz, 1978) are defined by

$$\begin{aligned} \text{AIC} &= -2\log\big(L(Y \mid \Theta_{mle})\big) + 2K, \\ \text{BIC} &= -2\log\big(L(Y \mid \Theta_{mle})\big) + K\log n, \end{aligned}$$

where Θ_{mle} is the sample point where the log-likelihood function is maximised after burn-in and K is the number of parameters, so that $K = 3(k-1)p-3+n+1$. It is well known that AIC is minimax-rate optimal in estimating the regression function (Barron et al., 1999; Yang, 2005), and BIC is consistent in model selection (Shao, 1997; Yang, 2005). Note that the model which has smaller DIC, WAIC, AIC or BIC provides a better model.

4 Application to forensic facial data

4.1 Data description

Facial features play an important role in forensic science including in criminal investigations where CCTV evidence is commonly used. A study was carried out into using face landmarks for identification, and was reported by Evison and Bruegge (2010). Clearly age and gender are expected

to be important covariates when describing the size and shape of the face landmark configurations, and so we develop some Bayesian regression models to explore the relationship. A set of 3D facial images was captured by a Geometrix FaceVision FV802 Series Biometric camera and then 30 anthropometric landmarks in 3D were selected by trained observers. The volunteers in the study were primarily scanned at the Magma Science Adventure Centre, Rotherham, UK. Evison and Bruegge (2010, Chapter 3) give full details of the project and provide discussion about the selection of the 30 landmarks. Many of the face landmark sets were recorded twice, either with different observers or the same observer. In total we have 3248 face landmark configurations from 1964 volunteers, in particular 956 faces from 627 females and 2292 faces from 1337 males. The landmark positions and descriptions are described in Table 1 following Evison and Bruegge (2010). Our main interest here involves investigating the relation between age and the size and shape of the faces for each gender.

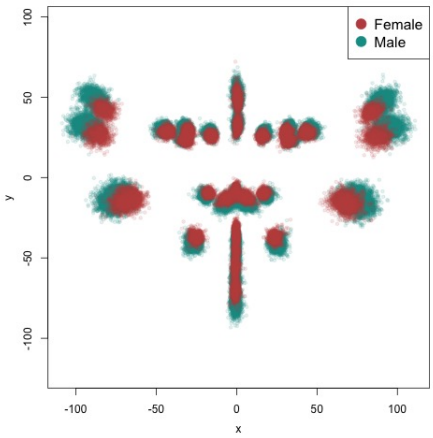
Table 1: Landmark information (Evison and Bruegge, 2010)

No.	Landmark	Label	No.	Landmark	Label
1	Glabella	g	16	Highest point of columella prime left	c' l
2	Sublabiale	sl	17	Highest point of columella prime right	c' r
3	Pogonion	pg	18	Labiale superius	ls
4	Endocanthion left	en l	19	Labiale inferius	li
5	Endocanthion right	en r	20	Stomion	sto
6	Exocanthion left	ex l	21	Cheilion left	ch l
7	Exocanthion right	ex r	22	Cheilion right	ch r
8	Center point of pupil left	p l	23	Superaurale left	sa l
9	Center point of pupil right	p r	24	Superaurale right	sa r
10	Palpebrale inferius left	pi l	25	Subaurale left	sba l
11	Palpebrale inferius right	pi r	26	Subaurale right	sba r
12	Subnasion	se	27	Postaurale left	pa l
13	Alare left	al l	28	Postaurale right	pa r
14	Pronasale	prn	29	Otobasion inferius left	obi l
15	Alare right	al r	30	Otobasion inferius right	obi r

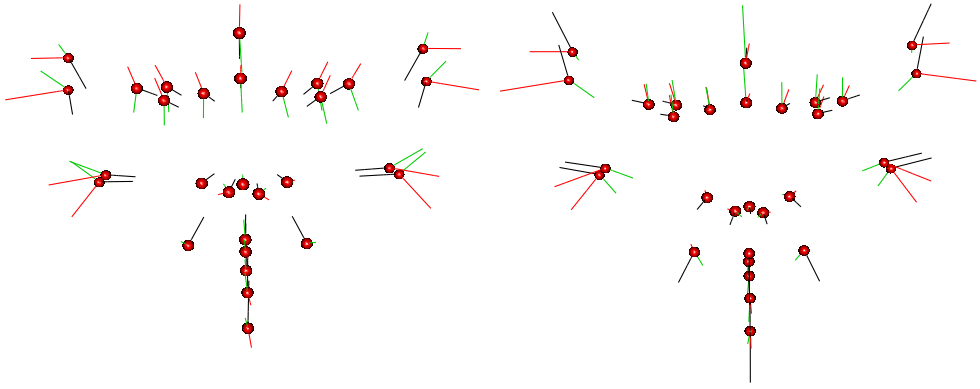
As would be expected on average male faces are larger and wider than female faces as shown in Figure 1 (a), where the main growth direction corresponds to the first shape principal component's direction indicated by black lines in Figure 1(b),(c). See Dryden and Mardia (2016, Section 7.7-7.8) for a summary of principal components analysis in shape and size-and-shape analysis, which has been implemented in `R` functions `procGPA()` and `shapepca()` in the package `shapes` (Dryden, 2017). In order to measure the size of the face landmark configuration, we use the centroid size of a configuration X given by

$$S(X) = \|HX\|,$$

where H is the Helmert submatrix (Dryden and Mardia, 2016, p.49) and $\|X\| = \sqrt{\text{trace}(X^\top X)}$. We see that the centroid size is closely related to the first size-and-shape principal component (PC) as seen in Figure 2 and the correlation coefficient between the centroid size and the first size-and-shape principal component score is -0.959 and 0.954 for female and male, respectively. Note that the signs of the PC loadings are arbitrary, and here PC1 and PC3 have different signs for females and males.



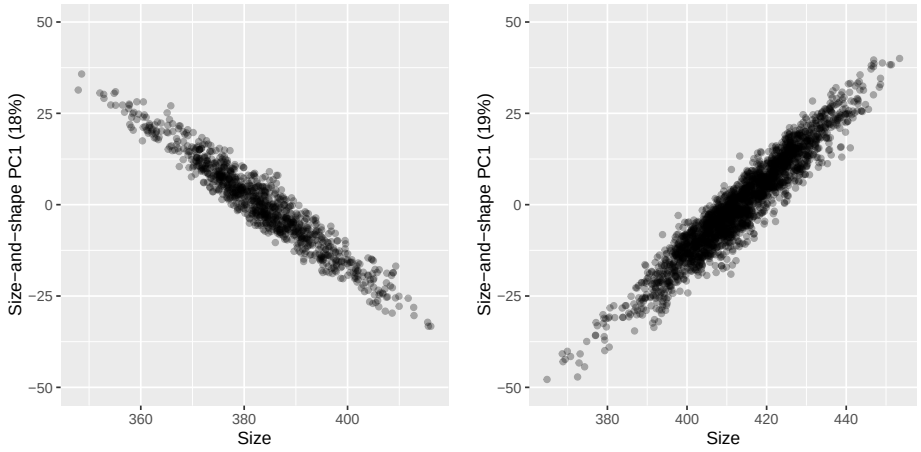
(a) Front view of size-and-shape configurations.



(b) Female

(c) Male

Figure 1: (a) Front view of size-and-shape configurations. (b), (c) Mean (red) and 3 PCs direction in $+3 \cdot$ sd (black: PC1, red: PC2, green: PC3).



(a) Female

(b) Male

Figure 2: Centroid size and size-and-shape PC1.

4.2 Models and implementation

Recall from Section 2.1 that the intercept matrix β_0 is lower triangular using an LQ decomposition for model identifiability, and the procedure is more stable if the landmarks in the first three positions are well separated. Hence in this application we re-ordered the landmarks as $(1, 3, 30, 2, 4, 5, ..., 28, 29)$. Note that the inference is invariant to a re-ordering of the landmarks, and so in theory such a re-ordering should make no difference in our modelling. However, in computational implementation it is best to avoid having the the first three landmarks too close together as otherwise some numerical instabilities can appear due to the standardisation via the LQ decomposition. The proposed re-ordering leads to stable results, which would in practice be equivalent to any other reordering with well separated landmarks in the first three positions.

For each gender we use the following three Helmertized models, where the Helmertizing takes care of the location invariance:

$$\begin{aligned}
\text{M1}: \quad Y_i^H &= \{\beta_0 + \beta_1 \mathtt{age}_i\}\,\Gamma_i + \varepsilon_i, \\
\text{M2}: \quad Y_i^H &= \left\{\beta_0 + \beta_1 \mathtt{age}_i + \beta_2 \mathtt{age}_i^2\right\}\Gamma_i + \varepsilon_i \\
\text{M3}: \quad Y_i^H &= \left\{\beta_0 + \beta_1 \mathtt{age}_i + \beta_2 \mathtt{age}_i^3\right\}\Gamma_i + \varepsilon_i,
\end{aligned}$$

where $Y_i^H = HY_i,\ i = 1, \ldots, n$. We consider a weakly informative conjugate prior κ so that $a = 0.001$ and $b = 1000$ (Spiegelhalter et al., 2003), and the hyperparameter F_0 in the prior distribution for the rotation parameters is taken as a 3×3 matrix of zeroes. For the MCMC algorithm we set the initial value of β to $\mathbf{0}$, $\Gamma_i, i = 1, \ldots, n$, to 3×3 identity matrices, and κ to a random draw from Gamma(a, b). The Gibbs samplers of Section 2.3 are used for updating $(\kappa, \beta, \theta_1, \theta_3)$ and in order to update θ_2 via the Metropolis-Hastings algorithm we use a normal distribution with standard deviation $\sigma_{\theta 2} = 0.3$ as the proposal distribution. To obtain a centred predicted face configuration we pre-multiply each fitted value $\widehat{Y}_i$ by C, for example for M2:

$$C\widehat{Y}_i = \left\{H^\top \widehat{\beta}_0 + H^\top \widehat{\beta}_1 \mathtt{age}_i + H^\top \widehat{\beta}_2 \mathtt{age}_i^2\right\} \widehat{\Gamma}_i,$$

where $C = I_k - \frac{1}{k} 1_k 1_k^\top$, I_k is the $k \times k$ identity matrix, 1_k is the column vector of k ones, $\widehat{\beta}_j = \frac{1}{T}\sum_{t=1}^{T} \beta_j^{(t)}$, $\widehat{\Gamma}_i = \frac{1}{T}\sum_{t=1}^{T} \Gamma_i^{(t)}$ are the arithmetic means of the MCMC sample of T iterations for β_j and Γ_i after burn-in, and in our faces application we have $k = 30$ landmarks.

4.3 Results

We run the MCMC chain for 200,000 iterations with 100,000 iterations of burn-in. The Metropolis-Hastings acceptance rate for θ_2 is around 3.72% for female and 3.83% for male data. The posterior variance for females is smaller than that for males as the posterior mean estimates for κ are larger than those for males in Table 2.

For the three models considered, model M2 is generally the best model for both female and male groups since M2 outperforms the others in terms of the model selection statistics DIC, WAIC and AIC in Table 3 (except that M1 has the smallest BIC for females). We now investigate the structure of the models of the fitted size-and-shapes of the face landmarks versus age and gender.

We display the results from the fitted regression models M1, M2, M3 for each gender in Figure 3, which shows the individual face landmark data registered by generalized Procrustes size-and-shape analysis (Dryden and Mardia, 2016, p.143) in light grey and viewed from the front projection of the

Table 2: Estimates for κ and acceptance rate for θ_2.

Model		Posterior mean κ	95% cred.int. κ	Acceptance rate θ_2
Female	M1	0.1124	(0.1115, 0.1133)	3.73%
	M2	0.1127	(0.1118, 0.1136)	3.72%
	M3	0.1127	(0.1117, 0.1136)	3.72%
Male	M1	0.0911	(0.0906, 0.0916)	3.85%
	M2	0.0925	(0.0920, 0.0930)	3.83%
	M3	0.0924	(0.0919, 0.0929)	3.83%

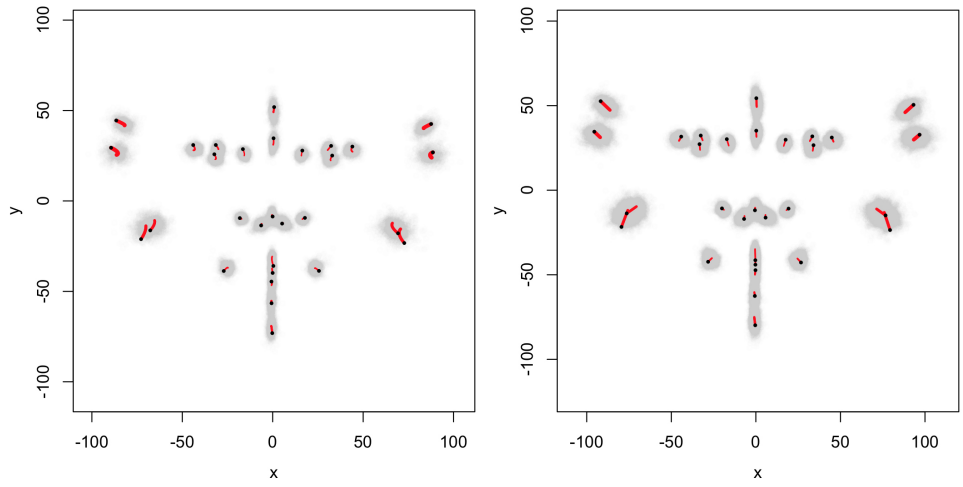
Table 3: Model selection statistics. Note that the best model for each line is indicated in bold.

Statistics	Female			Male		
	M1	M2	M3	M1	M2	M3
# of parameters	1128	1215	1215	2464	2551	2551
Maximum log-likelihood	-208753	**-208635**	-208649	-521561	**-520026**	-520187
DIC	420986	**420837**	420881	1050933	**1047960**	1048246
WAIC1	415056	**414899**	414944	1036202	**1033334**	1033605
WAIC2	416164	**416020**	416067	1038860	**1035994**	1036264
AIC	419763	**419699**	419727	1048050	**1045155**	1045475
BIC	**425248**	425607	425636	1062187	**1059790**	1060111

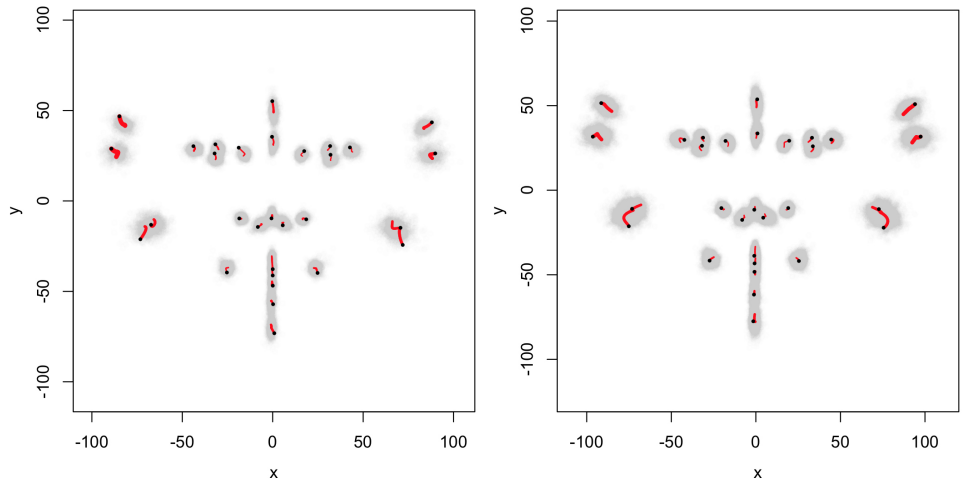
face. The fitted configurations can be arbitrarily rotated, and in order to compare the fitted configurations over age with the size-and-shapes of the Procrustes registered data, we apply ordinary Procrustes analysis to translate and rotate the fitted faces from the model (in red) onto the Procrustes mean size-and-shape of the data. The fitted faces are indicated by a red curved line from the fitted face at age 15 through to the fitted face at age 80 (which is identified with a black dot).

The fitted models for the females and males show important differences in Figure 3. In particular it is noticeable that the amount and direction of facial growth as age increases differ between females and males. For model M1, the males' fitted face linearly grows as age increases but the change is different for females as age increases. In some areas such as the eyes and ears, the face grows quicker later for females. The growth direction of the ears of females is relatively wider than that for males. The credible intervals for age for eight landmarks on the ears (landmarks 23 – 30), indicated by red line thickness, are relatively longer than the others for both the females and males. On the other hand, the credible intervals for the eight landmarks on the eyes (landmarks 4 – 11) are relatively shorter. For models M2 and M3, it is notable that the growth direction of the four landmarks on the bottom of the ears (landmarks 29, 30, 25 and 26) is different for females and males, where the females' ears grow wider than the males'. The results of models M2 and M3 are more similar to each other than M1 for both females and males. This observation can be inferred from Table 3 showing smaller differences in the model selection criteria for M2 versus M3 compared to M1 versus M2.

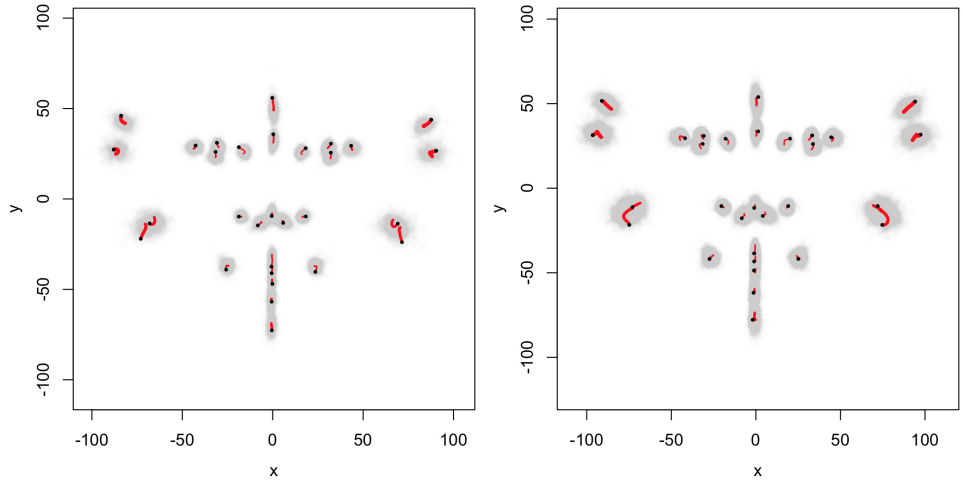
From now on we focus on the result of the model M2. Figure 4 shows magnified ears and lips. The main features that are apparent from Figure 4 are that as the faces become older the ears become larger and the lips become less full (i.e. thinner). Some of the curves have turning points where the



(a) Female M1 (b) Male M1

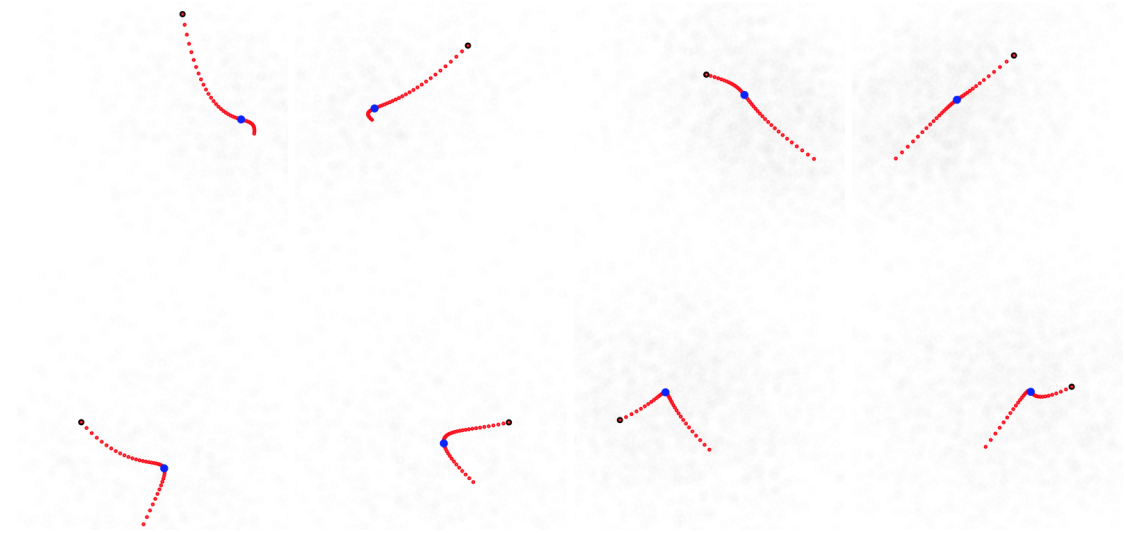


(c) Female M2 (d) Male M2

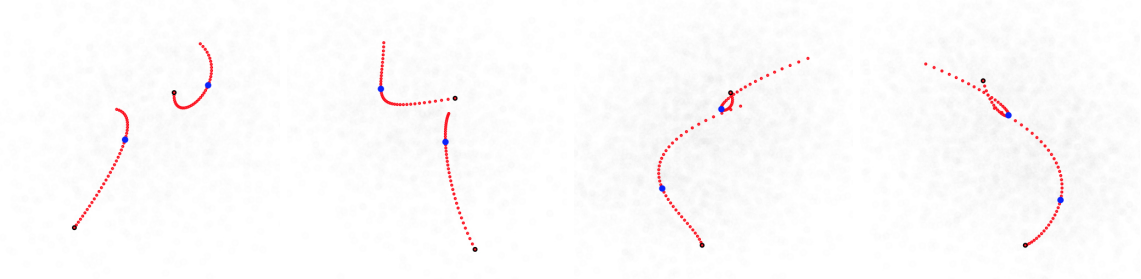


(e) Female M3 (f) Male M3

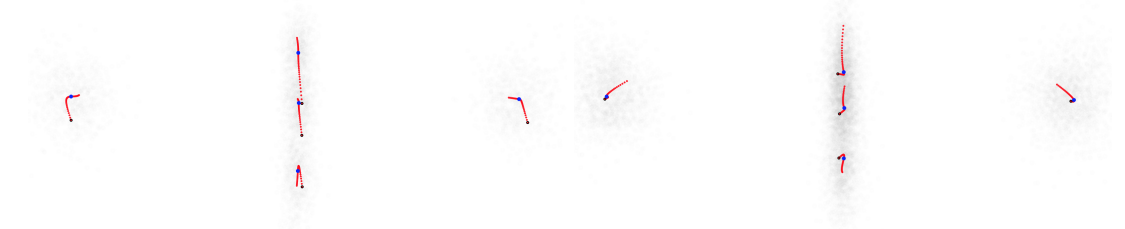
Figure 3: Front view. Light grey: Procrustes registered face data. Fitted values versus age: red lines. Credible interval for $\widehat{\beta}_1$: red lines' thickness. The thickness is proportional to the length of credible interval. Black dots: the fitted landmarks at age 80.



(a) Female: ear, top left. (b) Female: ear, top right. (c) Male: ear, top left. (d) Male: ear, top right.



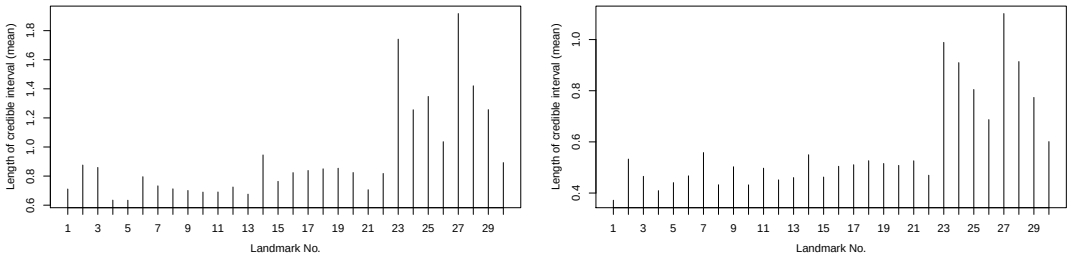
(e) Female: ear, bottom left. (f) Female: ear, bottom right. (g) Male: ear, bottom left. (h) Male: ear, bottom right.



(i) Female: lips. (j) Male: lips.

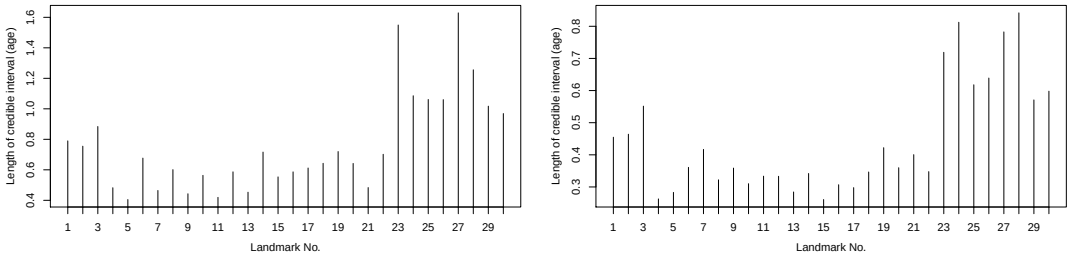
Figure 4: Ears and lip (M2).

behaviour is different before and after the turning point. The age at the turning point of the predicted curves can be different depending on the landmark position. We mark blue points to indicate age 37 for female and age 52 for male, and a black dot for age 80. Note that the predicted red points were obtained at equal age intervals. The speed of facial growth varies over age, for example for the top of the ears of females, (a) and (b), the upper parts of the top ears grow slowly for young women but those parts grow rapidly for older women. For the lower parts of the top of the ears, the speed of growth starts slowly and then becomes faster after age 37. For men in (c) and (d), a similar pattern to the lower parts of the top of the ears appears with the turning age 52. For the bottom of the ears and lips, (e) – (j) females and males show opposite results in growing speed, where females grow rapidly as age increases, however males grow slowly as age increases.



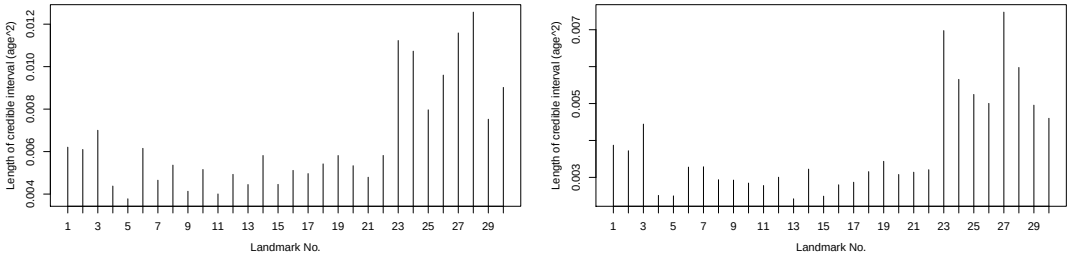
(a) Female: mean configuration.

(b) Male: mean configuration.



(c) Female: `age`.

(d) Male: `age`.



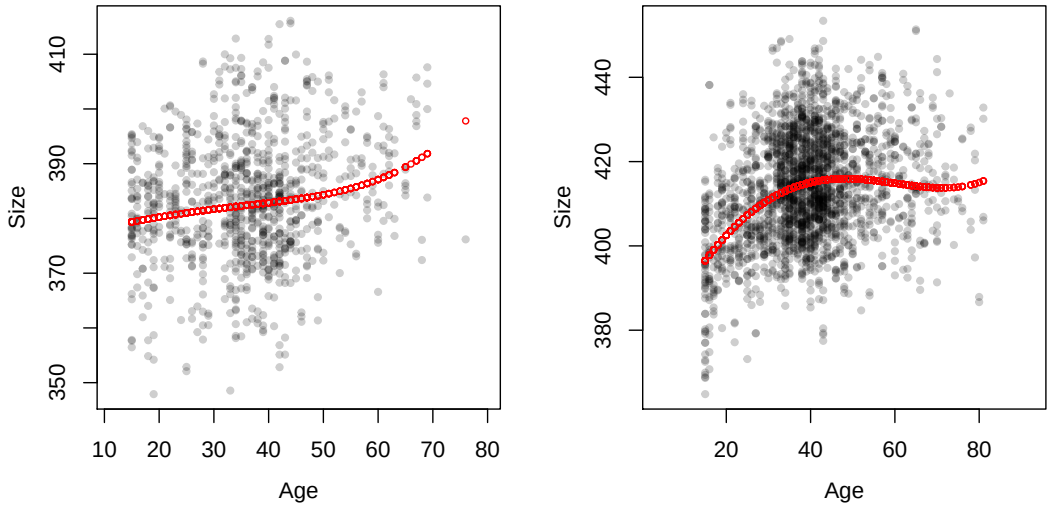
(e) Female: $\texttt{age}^2$.

(f) Male: $\texttt{age}^2$.

Figure 5: Length of credible interval (M2).

From Figure 5 (a) and (b) the outer parts of the face have more posterior variability which is shown in the length of the credible intervals for both ears (landmarks 23 to 30). In contrast near the eyes (landmarks 4 – 11), the lengths of credibility intervals are short. When faces grow as age increases for both females and males, the variability near both ears is larger as shown in (c), (d), (e) and (f).

In Figure 6 we see that the predicted centroid size for females is monotonically increasing. On the other hand male faces grow in size until age 40 then this stops, and so we can see important



(a) Female

(b) Male.

Figure 6: A scatter plot of age versus centroid size of raw configurations as dots and the centroid size of predicted configurations as red lines (M2).

differences here between the genders. Of course Figure 6 also illustrates the wide amount of individual variability in face data, and our model is just a first step in modelling average face shape. There is considerably more work required in modelling individual or sub-group face data, although our methodology provides a useful framework in which to develop these ideas.

Acklowedgements

This work was supported by the Engineering and Physical Sciences Research Council grant number EP/K022547/1 and Royal Society Wolfson Research Merit Award WM110140.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory*, pages 267–281. Budapest.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113:301–413.

Bowman, A. (2008). Statistics with a human face. *Significance*, 5(2):74–77.

Cheng, W., Dryden, I. L., and Huang, X. (2016). Bayesian registration of functions and curves. *Bayesian Anal.*, 11(2):447–475.

Cornea, E., Zhu, H., Kim, P., and Ibrahim, J. G. (2017). Regression models on Riemannian symmetric spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(2):463–482.

Davis, B., Bullitt, E., Fletcher, P., and Joshi, S. (2007). Population shape regression from random design data. In *IEEE 11th International Conference on Computer Vision*.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2018). Nonparametric rotations for sphere-sphere regression. *Journal of the American Statistical Association*. To appear.

Dryden, I. L. (2014). Shape and object data analysis [discussion of the paper by Marron and Alonso (2014)]. *Biom. J.*, 56(5):758–760.

Dryden, I. L. (2017). *Shapes: Statistical Shape Analysis*. R package version 1.2.3.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis, with Applications in R, 2nd edition*. Wiley, Chichester.

Evison, M. and Bruegge, R. V. (2010). *Computer-Aided Forensic Facial Comparison*. CRC Press.

Faraway, J. (2004). Human animation using nonparametric regression. *Journal of Computational and Graphical Statistics*, 13:537–553.

Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vis.*, 105(2):171–185.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. C. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imaging*, 23(8):995–1005.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton. third edition.

Goodall, C. R. and Lange, N. (1989). Growth curve models for correlated triangular shapes. In Berk, K. and Malone, L., editors, *Proceedings of the 21st Symposium on the Interface between Computing Science and Statistics*, pages 445–454. Interface Foundation, Fairfax Station.

Green, P. J. and Mardia, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93:235–254.

Gupta, A. and Nagar, D. (1999). *Matrix Variate Distributions*. Chapman & Hall/CRC.

Hinkle, J., Fletcher, P., and Joshi, S. (2014). Intrinsic polynomials for regression on Riemannian manifolds. *J. Math. Imaging Vision*, 50:32–52.

Hotz, T., Huckemann, S., Munk, A., Gaffrey, D., and Sloboda, B. (2010). Shape spaces for prealigned star-shaped objects—studying the growth of plants by principal components analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 59(1):127–143.

Huckemann, S., Hotz, T., and Munk, A. (2010). Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica*, 20(1):1–58.

Jung, S., Dryden, I. L., and Marron, J. S. (2012). Analysis of principal nested spheres. *Biometrika*, 99(3):551–568.

Jupp, P. E. and Kent, J. T. (1987). Fitting smooth paths to spherical data. *J. Roy. Statist. Soc. Ser. C*, 36(1):34–46.

Kendall, D. G. (1986). In discussion to 'size and shape spaces for landmark data in two dimensions' by F. L. Bookstein. *Statistical Science*, 1:222–226.

Kendall, D. G. (1989). A survey of the statistical theory of shape (with discussion). *Statistical Science*, 4:87–120.

Kendall, D. G., Barden, D., Carne, T. K., and Le, H. (1999). *Shape and Shape Theory*. Wiley, Chichester.

Kenobi, K., Dryden, I. L., and Le, H. (2010). Shape curves and geodesic modelling. *Biometrika*, 97(3):567–584.

Kent, J. T., Mardia, K. V., Morris, R. J., and Aykroyd, R. G. (2001). Functional models of growth for landmark data. In Mardia, K. V. and Aykroyd, R. G., editors, *Proceedings in Functional and Spatial Data Analysis, LASR2001.*, pages 109–115. University of Leeds.

Kume, A., Dryden, I. L., and Le, H. (2007). Shape-space smoothing splines for planar landmark data. *Biometrika*, 94(3):513–528.

Landau, L. D. and Lifschitz, E. M. (1976). *Mechanics, 3rd edition*. Pergamon Press, Oxford.

Le, H. and Kume, A. (2000). The Fréchet mean shape and the shape of the means. *Adv. in Appl. Probab.*, 32(1):101–113.

Machado, L. and Leite, F. (2006). Fitting smooth paths on Riemannian manifolds. *International Journal of Applied Mathematics and Statistics*, 4:25–53.

Mardia, K. (1975). Statistics of directional data. *Journal of the Royal Statistical Society, Series B*, 37:349–393.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley, Chichester.

Piras, P., Evangelista, A., Gabriele, S., Nardinocchi, P., Teresi, L., Torromeo, C., Schiariti, M., Varano, V., and Puddu, P. E. (2014). 4D-analysis of left ventricular heart cycle using Procrustes motion analysis. *PLOS One*, 9(4):e94673.

Presnell, B., Morrison, S., and Littell, R. (1998). Projected multivariate linear models for directional data. *Journal of the American Statistical Association*, 93:1068–1077.

Rosenthal, M., Wu, W., Klassen, E., and Srivastava, A. (2014). Spherical regression models using projective linear transformations. *Journal of the American Statistical Association*, 109(508):1615–1624.

Rosenthal, M., Wu, W., Klassen, E., and Srivastava, A. (2017). Nonparametric spherical regression using diffeomorphic mappings. *ArXiv e-prints*. 1702.00823.

Samir, C., Absil, P.-A., Srivastava, A., and Klassen, E. (2012). A gradient-descent method for curve fitting on Riemannian manifolds. *Foundations of Computational Mathematics*, 12(1):49–73.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264.

Shi, X., Styner, M., Lieberman, J., Ibrahim, J., Lin, W., and Zhu, H. (2009). Intrinsic regression models for manifold-valued data. In *In Proc Int. Conf. Medical Image Computing and Computer-assisted Intervention*, pages 192–199. London, Sept. 20th-24th (eds G.-Z. Yang, D.J. Hawkes, D.Rueckert, A. Noble and C. Taylor). Berlin: Springer.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., and Lunn, D. (1994, 2003). Bugs: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England.

Su, J., Dryden, I. L., Klassen, E., Le, H., and Srivastava, A. (2012). Fitting smoothing splines to time-indexed, noisy points on nonlinear manifolds. *Image and Vision Computing*, 30:428–442.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92:937–950.

Yuan, Y., Zhu, H., Lin, W., and Marron, J. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society, Series B*, 74:697–719.

Zhu, H., Chen, Y., Ibrahim, J., Li, Y., Hall, C., and Lin, W. (2009). Intrinsic regression models for positive definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association*, 104:1203–1212.